Projet Supervised Machine Learning: Heart Failure Prediction

SEREME Ousmane
ZONGO Celse Darius Pegdwende
ZONGO Rasmane

Sept 2021

Contents

1 Introduction

Ce projet est initié dans le cadre de notre Master Fouilles de Données et Intelligence Artificiel. Il nous permettra de renforcer nos connaissance en Machine Learning.

Problème choisi : Nous avons choisi pour ce projet de travailler sur un problème de classification. En effet notre sujet traite de la prédiction d'un arrêt cardiaque entraînant le décès du patient.

contexte : les maladies cardio-vasculaires sont la principale cause de mortalité dans le monde actuellement (environ 17 million de mort en 2019). Notre travail consistera à définir un modèle de prédiction des décès survenu après un arrêt cardiaque. Ce modèle pourrait servir à conseiller les patients et la populations su les critères qui peuvent grandement influencer la mortalité de ces maladies.

2 la base de données et logiciels utilisées

la base de donnée utilisée pour ce travail est de Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020). Elle contient 13 paramètres et 299 entrées, que nous avons trouve sur Kaggle.

Pour cette étude nos avons utilisée python avec la bibliothèque Pyspark. Nous avons utiliser l'editeur de texte pyspark.

3 analyse exploratrices

Les sujets répertorier dans cette base de donnée, sont de personnes entre 40 et 95 ans. Ils sont à majorité de sexe masculin. Les données ont déjà été transforme en données numériques (boolean, et categorical data en numerique). Ceci permettra d'utiliser directement la régression logistique sans besoin d'utiliser un long pipeline pour normaliser les données. De plus les données sont faiblement corrélées, donc chaque paramètre sera important dans la construction de nos modelès.

4 modèles implémentés

En rappel, ce problème est un problème de classification binomiale. En effet le résultat c'est la mort du patient ou non. Nous avons alors utiliser les algorithmes suivant:

- **régression logistique:** c'est un algorithme permettant d'étudier les relations entre un ensemble de variables Xi et une variable Y. Il s'agit d'un modèle linéaire . Un modèle de régression logistique permet aussi de

prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression.

- **arbre de choix :** repose sur des arbres de décisions, c'est-à-dire des séries de questions à choix multiples qui mène à la décision finale. L'algorithme de machine Learning va permettre par itération de définir les probabilités d'arriver à une décision. Cela permet d'optimiser le chemin à suivre pour arriver au bon résultat.
- **Foret aléatoire :** est une méthode d'ensemble. Elle est compose de plusieurs arbre décisionnel.
- **Arbre dégrade:** est également une méthode d'ensemble composes d'arbres décisionnels. Il peut être utile quand l'arbre décisionnel a de faible résultats dans l'apprentissage. Dans ces cas il peut même être plus performant que la forêt aléatoire.

5 Évaluations

Dans le tableau ci dessous nous avons les mésures des différents algorithme utilisés.

mesures	régression logistique	arbre de choix	foret aléatoire	arbre dégrade
Accuracy	0.78	0.83	0.86	0.80
sensibilité	0.47	0.58	0.63	0.58
spécificité	0.87	0.90	0.93	0.87
Précision	0.50	0.61	0.71	0.55
AUC	0.81	0.78	0.86	0.79

De la nous voyons que le meilleur modèle est celui de la foret aléatoire.
Au besoin nous pouvons utiliser l'hypertuning des paramètres pour améliorer les mesures de la foret aléatoire.

6 conclusion

Pour la principal cause de mortalité,(17 millions de personnes par ans) la base de donnée n'est pas assez fourni (rien que 300). Par ailleurs n'ayant pas d'informations dur le processus de collecte de données. On pourrait craindre un biais dans la collecte des données .
Il serait intéressant d'avoir beaucoup plus de sujets, augmenter les paramètres médicaux (avec une base de données plus représentative, certains paramètres non répertoriés pourrait êtres utiles)